DOCUMENT RESUME

ED 217 062 TM 820 277

AUTHOR Easton, John Q.; Washington, Elois D.
TITLE The Effects of Functional Level Testing on Five New Standardized Reading Achievement Tests.
PUB DATE Mar 82
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; Pilot Projects; *Reading Achievement; *Scores; *Standardized Tests; Testing Programs; *Test Reliability; *Test Selection; Test Validity
IDENTIFIERS California Achievement Tests; Iowa Tests of Basic Skills; Metropolitan Achievement Tests; *Out of Level Testing; Sequential Tests of Educational Progress; SRA Achievement Series; Testing Conditions; *Test Levels

ABSTRACT

The effects of students taking different levels of the same standardized achievement test were assessed by administering two levels of the same test to each student. The functional level of the test was taken by all students. The second level of testing was randomly assigned at the adjacent higher or lower level of the test. Functional level testing is the method of using students' instructional level rather than age or grade level to assign achievement test levels. Test reliability and stability of scores from one test level to another were studied. Content validity was also a major factor in determining which test was suitable for use. Testing appeared to be more reliable in high achievement schools than in low achievement schools. Three of the five tests evaluated were determined to be suitable for Chicago elementary schools. Two of the tests were found to be less suitable because it was difficult to determine an accurate means of student placement. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED217062

The Effects of Functional Level Testing on Five New Standardized Reading Achievement Tests

John Q. Easton
City Colleges of Chicago

Elois D. Washington
Chicago Board of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Q. Easton
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 277

Paper presented at the annual meeting of the
American Educational Research Association
New York City, March 1982

Acknowledgements

This project was conceived, designed, executed, and completed by the cooperation of numerous individuals, departments, and test companies. The need for this study was expressed by many officials at the Chicago Board of Education and this need was articulated at the Department of Research and Evaluation with the assistance of a committee of district superintendents, principals, and other testing specialists in the school system. The design of the study was prepared by Irving Brauer, John Easton, and Elois Washington. Under the direction of Elois Washington, the pilot study was executed in sixty schools by principals, counselors, and entire teaching staffs. Throughout the entire process of this study the representatives of the five test publishers have been highly professional, helpful, and enthusiastic. At one time or another, all of the staff at Research and Evaluation was involved in this study. Outstanding among them are Elmer Casey, Carole Perlman, Bill Rice, Frank Ward, and Lynne Williams.

Note

The first author of this paper was a paid consultant for the Department of Research and Evaluation at the Chicago Board of Education when this study was conducted. The recommendations and conclusions of this paper do not necessarily reflect the position of the Chicago Board of Education.

The Effects of Functional Level Testing on Five New Standardized Reading Achievement Tests

The Department of Research and Evaluation of the Chicago Board of Education designed and implemented a large scale trial testing program of five commercially prepared standardized achievement tests in preparation for selecting a new standardized achievement test for the citywide elementary school testing program. The objectives and purposes, the procedures, and the results of the pilot testing program are described in this paper.

Functional level testing, also called out-of-level testing, is the practice of using students' instructional level rather than grade-level or age-level to assign students to achievement test levels. In Chicago, at the time of this study, the instructional level was the level on the Continuous Progress/Mastery Learning reading program that determined students' placement on the Iowa Tests of Basic Skills, and not grade or age. The educational reasoning for functional level testing is that students should be tested on masterials that resemble what they have encountered in classroom instruction, and not material that is advanced beyond their current instruction. A practical consideration favoring functional level testing in Chicago (see Wick and Ward, 1977) is that many students performed at or below chance level scores when the tests were administered on grade level, so that the reliability of such scores was highly questionable.

One of the major purposes of this pilot testing program was to study various consequences of functional level testing, and to determine the ability of several different tests to perform reliably under various functional level testing conditions. This condition of reliability in functional testing conditions was added to the usual list of criteria that are studied when a school system contemplates a new testing program. Among

the more usual criteria are how well the test objectives correlate with the school system's instructional objectives, how easily the test can be administered and scored, whether the test items are free from racial and sexual bias, and how much the test costs and what kind of services the test publisher can provide to the school district. These criteria have been examined by various Board of Education committees and are reported in internal documents.

This study was designed by staff members at the Department of Research and Evaluation in order to study the effects of on- and off-level testing (functional level testing) on students' scores, and on the reliability of test scores. According to five test publishers involved in this trial program, and according to our research, no such large scale empirical study of different achievement tests has been undertaken by any school district in its test selection process.

Although out-of-level testing practices have been adopted by many school districts in the past decade, the effects of out-of-level testing on students' scores and on test reliability are not well known. In an early study, Ayrer and McNamara (1973) concluded that out-of-level testing decreased guessing-level scores on the Iowa Tests of Basic Skills, but that students got different scores from one level of the test to another. Long, Schaffran, and Kellogg (1977) found that grade equivalent scores on the Gates-MacGinities reading comprehension and vocabulary scales increased when second and third grade students took a test level at least one level lower than their grade but that scores decreased when fourth graders took the second or third grade level instead of the fourth grade level. Agrawal (1978), in a study conducted at the Department of Research and Evaluation in Chicago, demonstrated that higher levels of the Iowa Tests of Basic

Skills reading comprehension test produced higher grade equivalent scores than lower levels of the same test. In one final study, Yoshida (1976) reported that chance level scores decreased and reliability coefficients increased when EMH students were assigned to low levels of the Metropolitan Achievement Test by teacher judgment. This current study is a comprehensive study of the effects of functional level testing on both reliability and students' scores.

Because of the time involved, and other practical considerations, the reading achievement subtest was the only portion of the test batteries that was administered in this study. The five tests in the study are the California Achievement Test, the Iowa Tests of Basic Skills, the Metropolitan Achievement Test, and Scientific Research Associates Achievement Series, and the Sequential Test of Educational Progress. These tests are thoroughly described and compared in a recent article by Iwanicki (1980).

Design of the Study

Since the study's primary purpose was to determine the effects of students' taking different levels of the same test, every student in the study took two levels of the same test. (One level, taken by all students, was the "functional" level of the test, and the other level was either the adjacent higher level or the adjacent lower level of the test. Two classes of exceptions to this rule will be explained later.) At the time of the study, Chicago did not have grade levels in the elementary schools, so we relied on placement in the Chicago Continuous Progress Reading Program (CP) to define what level of each test was functional (comparable to the instructional level of each student). After a grade was assigned for the continuous progress levels, we used the publishers' guideline to assign the correct level of each test to students. Table 1 contains the

level of each of the five tests in this study that came to be called the "functional level" for each continuous progress level. As the table shows, the tests have different "widths"--that is, for some tests there is a different level for each grade or CP level, while other tests are designed to be appropriate for a wider range of students.

Table 1

Functional Testing Levels

CP Level	CAT	ITBS	Metropolitan	SRA	STEP
A or B	10	5	Primer	A	Circus B
C or D	11	7	Primary 1	A	Circus C
E or F	12	8	Primary 2	B	Circus D
G or H	13	9	Elementary	C	STEP E
J	14	10	Elementary	D	STEP F
K	15	11	Intermediate	E	STEP G
L	16	12	Intermediate	E	STEP H
M	17	13	Advanced I	F	STEP I
N	18	14	Advanced I	F	STEP I

As we said earlier, every student in the study took the functional level of a test (as defined by the criteria in Table 1); half of these students also took the adjacent lower level of the test, and the second half took the adjacent higher level of the test. The students were randomly assigned to either the higher or lower level of the test within classroom by the test administrator. For example, if we had 20 J students in one of the schools in the study assigned to the ITBS, all 20 would take Level 10, and a random half (perhaps every other student, or every other row of students) would take Level 9 and the remaining half would take Level 11. The order of the testing was counterbalanced by classroom within school, so that in every school in the study half of the classrooms took the functional test level in the morning and the off-level tests in the afternoon. In another school in the study, there were J students taking

the off-level tests in the morning and Level 10 in the afternoon. The purpose of the counterbalancing was to eliminate any effects that practice, or on the other hand, fatigue would have on test scores.

There were sixty schools that participated in this study. We made five sets of schools with twelve schools in each set, and assigned each set of schools to one of the five achievement tests. Each set of twelve schools contains three schools from each of four achievement-level quartiles of Chicago's elementary schools, so that each set has an equivalent number of schools throughout the achievement range. In addition to being nearly equivalent in terms of the percent of students from poverty level homes, the sets of schools have nearly equal overall reading indices (scores prepared by the Department of Research and Evaluation to represent the reading achievement level of the entire school) and contain schools of about the same average size. The sets of schools for each of the five tests are compared more fully in Appendix A of this report, the "Description of Schools in the Study."

It was necessary to control age in the design of the study because many students in low achievement schools in Chicago are older than students of the same Continuous Progress levels in high achievement schools. In order to systematize these naturally occurring age differences, we imposed age restrictions on which students could enter the sample. The ages of the students in the 25% highest achieving schools (Q4) correspond to the average age level "equivalent" to their CP levels; that is, in Q4 schools we tested 6-year old CDs (because CD translated to first grade), 7-year old EFs (EF translated to second grade), 8-year old GHs (GH = third grade), 9 year-old (J = 4th grade), 10 year-old Ks (K = fifth grade), 11 year-old Ls (L = sixth grade), 12 year-olds Ms (M = seventh grade), and 13-year old Ns (N = eighth grade). The students from Q3 and Q2 were all one year

older than students of the same CP levels in the Q4 schools. The students in A1 schools were all two years older than the students of the same CP levels in the Q4 schools. This pattern is shown in Table 2.

Table 2

Age Levels Chosen from Schools of Different Achievement Levels

CP Level	Q4 Schools	Q3 Schools	Q2 Schools	Q1 Schools
A or B	-	6	6	7
C or D	6	7	7	8
E or F	7	8	8	9
G or H	8	9	9	10
J	9	10	10	11
K	10	11	11	12
L	11	12	12	13
M	12	13	13	14
N	13	13	14	-

The functional level tests for the students in Q4 schools are equal to grade-level tests, the functional level tests for students in Q3 and Q2 schools are one year below grade-level tests, and the functional test levels of students from Q1 schools are two years below grade-level tests. Because the Q1 students were taking tests two years behind their grade (age) levels, we did not give them a lower level of the test. All Q1 students took the functional test level and the adjacent higher level only. (This is one exception to the design of the study. The other exception occurred when a test was too wide to permit adequate testing of the same student group from three different test levels. For example, J students who were tested on the Metropolitan test were not tested on a lower level, since the next lower level, Primary 2, was clearly inappropriate for them. All J students took the Elementary level and the Intermediate level of the Metropolitan.)

Summary of the Design of the Study

Every student in the study took two levels of a reading test: the level corresponding to the student's instructional reading level (the functional test level), and either the adjacent higher or lower test level. Each of the five standardized reading tests in the study was administered in twelve schools, which were chosen to represent the city's schools as a whole, and to be nearly equivalent to the sets of twelve schools assigned to the other tests.

Results of the Study[1]

There are two major questions that we address here and attempt to answer with the results of this study. The first question is, "How reliable is each of the tests when it is administered under the conditions we imposed?" The second question, equally important, is, "How much difference does it make, in terms of scores, which level of a test is administered to students?"

Before we go on to these issues, we will examine the extent and size of this study. Table 3 on the following page contains the numbers of students whose test scores were used in the analyses reported here. These numbers are not strictly equal to the number of students who were actually tested, since not all scores were included in the analyses (for example, the lowest level of Test E in its present form is not amenable to statistical analysis). About 2,000 students are included in the analysis of each test, with the exception of Test E, where there are somewhat fewer students (about 1,750).

Reliability

The concept of reliability is a crucial one to test theory. The

[1]For the remainder of this paper the five achievement tests in the study are referred to as Test A, Test B, Test C, Test D, and Test E.

Table 3

Number of Scores Analyzed in this Study

Age	Test Level	A	B	C	D	E
6	Functional	246	250	---*	247*	33*
	Lower or Higher	259	252	307	49	272
7	Functional	226	250	307	298	240
	Lower or Higher	213	254	314	237	198
8	Functional	306	258	268	243	277
	Lower or Higher	303	255	278	238	269
9	Functional	249	329	283	253	356
	Lower or Higher	238	290	285	247	354
10	Functional	227	211	217	328	302
	Lower or Higher	209	229	240	344	302
11	Functional	237	239	---*	206	161
	Lower or Higher	236	236	285	205	161
12	Functional	250	196	264	185	126
	Lower or Higher	251	197	254	182	128
13 or 14	Functional	336	293	352*	366	250*
	Lower or Higher	337	323	129	370	103
TOTAL	Functional	2,077	2,026	1,691*	2,126	1,745
	Lower or Higher	2,046	2.036	2,092	1,872	1,787

*When a major discrepancy occurs between the number of students on the functional level and on the lower or higher level it is for one of three reasons: 1)there is no higher or lower test level, 2)the scores are not reported in comparable fashion on the higher or lower level or 3) there is an inaccuracy or error in reporting.

technical meaning of the term is highly similar to the general, popular meaning--what can be relied on or is dependable is reliable. Scores from a test that is free from error and accurately represents whatever attribute is being scrutinized are reliable scores. A reliability coefficient estimates the portion of variation in a set of obtained scores that is due to actual variation in the attribute (reading achievement) and that due to error of measurement (Thorndike, 1971, p. 374). A reliability coefficient of .85 means that 85% of the variation in a set of obtained test scores is due to real differences in reading achievement, and the remaining 15% of the variation in scores is due to testing error. What counts as testing error varies from one means of estimating reliability to another. The Kuder-Richardson formula 20 (KR20) that we used in this study as a reliability coefficient estimates the consistency of test items at one given testing time. The coefficient estimates how consistently the test items measure reading achievement. Item idiosyncracies and idiosyncracies of the students in regard to particular test items are the error components in the KR20 reliability coefficent.

The test publishers in this study computed KR20s for every subgroup of students tested, as identified by age level and Continuous Progress level. The analysis produced 56 coefficients for the Test B, 54 for Test A, and 45 each for the Test D and Test E. (Reliability coefficents for Test C were not available when this paper was prepared.) In order to make meaningful comparisons among all of these reliability coefficents, they have been summarized into different sets of coefficients that are reported in Table 4 on the following page. In this table, the coefficients are broken down by the quartiles of the schools, then by whether the level of the test corresponded to the students' functional level or was the

Table 4

Median* KR20 Reliability Coefficients

Q4 Schools	Test A	Test B	Test C	Test D
Ss taking higher level tests[1]	.86	.92	.82	.87
Ss taking functional level[2]	.81	.88	.76	.84
Ss taking lower level test[3]	.75	.88	.60	.92
Q3 and Q2 Schools				
Ss taking higher level tests[2]	.81	.76	.82	.87
Ss taking functional level[3]	.81	.84	.79	.86
Ss taking lower level[4]	.81	.82	.61	.84
Q1 Schools				
Ss taking higher level[3]	.79	.77	.81	.83
Ss taking functional level[4]	.80	.80	.68	.85

[1]Test level one year above age level

[2]Test level equal to age level

[3]Test level one year lower than age level

[4]Test level two years lower than age level

*The median is based on all test levels given.

higher or lower test level. Footnotes in the table indicate how a test level corresponded to students' age levels.

The first generalization that comes apparent from inspecting this table is that testing is more reliable in the high achievement schools that in the low achievement schools (with an exception). The median coefficients (the median is based on all different age groups tested within each group) range between .75 and .86 on Test A, between .80 and .92 on Test B, between .60 and .82 on Test D, and between .83 and .92 on Test E. Given these ranges of median coefficents, the highest reliabilities occur on Test E, the next highest on Test B, next on Test A, and lowest on Test D. Although the differences are not great from one test to another, and the ranking is not constant for each subgroup of students, our data do show Test D to have nearly always the highest reliability coefficients, Test B and Test A to be close for second highest, and Test C to have the lowest reliability coefficients.

The next critical question to examine is whether tests are more reliable when they are assigned according to curricular level rather than according to age level or under other conditions. We are able to make direct comparisons between functional level testing and testing on either the higher or lower level in Table 4. Functional level testing appears to give slightly more reliable test results than does either the higher or lower level testing, especially in the low achievement schools. If we look at the tests one by one, we see that Test E, the most reliable test overall, does not vary grealy in reliability from one level to another, except in the Q4 schools, where for an unknown reason, both the higher and lower test levels are more reliable than the functional test level.

Test A is most reliable in the higher level in Q4 schools, equally reliable at all levels in the Q3 and Q2 schools, and slightly more reliable at the functional level in the Q1 schools than at the higher level. Test B is also more reliable at the higher level in the Q4 schools, most reliable at the functional level in the Q3 and Q2 schools and more reliable at the functional level in the Q1 schools. Test D is most reliable on the higher levels in the Q4 schools, most reliable on the higher levels in the Q3 and Q2 schools, and more reliable on the higher level in the Q1 schools. In Test E, the most reliable of the tests, the level makes little difference in reliability, whereas in Test D, the least reliable of the tests, the level makes a great difference. The two intermediate tests are most reliable at the higher level in Q4 schools, and at the functional level in all other schools.

We have been discussing the difference between functional level testing and test levels that are higher or lower, without having considered if the higher or lower test level is an age-level test. In the Q3 and Q2 schools, we are able to see the direct difference between functional level testing and age level testing. The Q3 and Q2 students who took the higher level test were taking a test equivalent to their age level. Test E is slightly more reliable on the age level test (.87 vs. .86), Test A equally reliable, Test B more reliable on the functional level than on the age leve, and Test D more reliable on the age level test than on the functional level test. In general, across all of these tests, there is little difference in median reliability for the age-level tests versus the functional level test (one year behind) in the middle 50% of Chicago schools. As for the Q1 schools (two years behind), functional level testing is somewhat more reliable than testing on a higher level in three of the four tests.

These findings may be summarized in the following statements: Test E is the most reliable of the tests, and on Test E there is the least amount of difference in reliability from one level to another. Test D is the least reliable, and the reliabilities change greatly according to which test level is administered to the students. Of the two other tests, Test A and Test B, testing is more reliable on the functional test levels than on the higher test levels (which correspond more closely to age level testing) in Q3, Q2 and Q1 schools but in Q4 schools testing is more reliable on higher level tests -- where we note that the higher level test is not equal to the age level test, but is one level above the age level test.

Several factors contribute to differences in reliability coefficients from one test to another. Among these factors are the average difficulty of the items on the test, the variability in the difficulties of the items, and the number of items on a test. When a test is too difficult for students and the responses are random or are guesses, then the test reliability suffers since the scores are due to error rather than to students' ability or achievement. Each of these three factors mentioned here played some role in determining the reliability coefficients that we obtained in this study.

Score changes

In this second section of the results the study we are examining various approaches to the question, "What is the effect of assigning different levels of a test on students' scores?" In a well constructed and well scaled test, there are few differences in scores from one level of the test to another, except among students for whom one of the test levels is much too easy or difficult. Regardless of the policy adopted

by the Board of Education concerning how students are to be assigned to test levels, there will always be students who are placed in an inappropriate test level. How much will this mean to the students' scores?

One of the most common scores that is used in standardized testing is a scale score that spans all levels of a particular test. Each of the five tests in our study has a scale score with a different range of possible scores. (Although grade equivalent scores have nearly the same range of scores from one test to another, we did not use these scores in this analysis.) Because scale scores have different values and ranges of scores, the difference between any two scale scores on one test has no comparability to the difference between two scale scores on a different test. In order to overcome this limitation (since, in effect, we want to know which test changes scores the least from one level to the next) we have developed a special standard scores (a z-score) to express the difference between scale scores on the functional test level and either the lower or higher test level that can be compared from one test to another. The z-scores, as we use them here, indicate the relative amount of difference between mean scores on two different test levels. These z-scores assume that the students in the comparison groups are random samples of the populations (students in the functional test level group), and provide an answer to the question, "What is the likelihood that the comparison group of scores is a sample from the same distribution of scores as the population scores?" Each z-score was calculated by subtracting the comparison mean from the functional test level mean and dividing the differences by the standard error of the functional test level distribution. A z-score of one means that the difference between the functional level and the comparison level is equal to the standard error of the functional test level.

We calculated a z-score for the difference between each lower level test and its functional test, and between each higher level tests and its functional test level for every subgroup of students (Continuous Progress level by age) in this study. An extensive summary of these statistics appears on Table 5 on the following two pages. In this table, for the purposes of summary, the students are broken down into three groupings -- the students from Q4 schools whose functional test levels are equal to their age level test (group I), students from Q3 and Q2 schools whose functional test levels are one year below their age-level tests (group II), and students from Q1 schools whose functional test levels are two years below their age-level tests (group III). There are difference measures reported under each of these groupings. For groups I and II there are separate categories for the difference between higher levels and functional levels, between lower levels and functional levels, and between the higher levels and the functional levels. (In the final case t-scores rather than z-scores are reported.) In group III there is only one category -- for the difference between the higher level scores and the functional scores, since the students in Q1 schools were only tested up. The table contains median z- and t-scores rather than a single statistic for each age group within groups I, II, and III.

Table 5 also contains other information. First of all there is the number of comparisons on which the median is based -- that is the number of age levels where there was an up or down testing. There are fewer on Test C and Test D because those two tests have wider levels and there were fewer times when a group was tested on the functional level and the up and down level. The second line on the table has the median z- or t-score, and the third line has the percent of z- or t-scores that are

Table 5

SUMMARY OF COMPARISONS OF SCORES ON
DIFFERENT TEST LEVELS

The following four values are reported for each group of students for each test:

1. The number of age levels where a comparison was made.
2. The median absolute value of the z-scores or t-scores used for comparing scores from one level to another (z-scores were used to compare adjacent test levels, t-scores to compare non-adjacent levels).
3. Percent of comparison scores less than -2 or greater than 2.
4. Percent of comparison scores with negative values.

I. Students whose functional test level is equal to their age level test.

A. Difference between higher level score and functional level scores.

	Test A	Test B	Test C	Test D	Test E
1.	7	7	4	7	6
2.	1.04	2.08	1.40	1.84	0.57
3.	28.6%	57.1%	50.0%	28.6%	16.7%
4.	0.0%	14.3%	50.0%	28.6%	0.0%

B. Difference between lower level score and functional level scores.

	Test A	Test B	Test C	Test D	Test E
1.	7	8	5	5	6
2.	3.05	3.64	1.17	3.70	1.25
3.	57.1%	75.0%	20.0%	50.0%	33.3%
4.	100.0%	100.0%	40.0%	75.0%	67.0%

C. Difference between high level score and lower level score.

	Test A	Test B	Test C	Test D	Test E
1.	6	7	2	3	5
2.	3.14	3.07	1.97	3.31	0.94
3.	67.0%	86.0%	50.0%	100.0%	20.0%
4.	0 %	0 %	0 %	33.0%	0 %

Table 5 Continued

II. Students whose functional test level is one year lower than their age-level test.

A. Differences between higher level scores and functional level scores.

	Test A	Test B	Test C	Test D	Test E
1.	8	7	5	6	6
2.	2.40	5.11	2.84	1.55	2.15
3.	63.0%	86.0%	80.0%	50.0%	50.0%
4.	38.0%	14.0%	80.0%	67.0%	50.0%

B. Difference between lower level scores and functional level scores.

	Test A	Test B	Test C	Test D	Test E
1.	7	8	5	4	6
2.	2.81	1.98	1.49	2.33	2.19
3.	57.0%	50.0%	40.0%	75.0%	67.0%
4.	86.0%	63.0%	20.0%	75.0%	33.0%

C. Difference between higher level scores and lower level scores.

	Test A	Test B	Test C	Test D	Test E
1.	7	7	2	3	5
2.	2.03	4.84	1.61	1.47	2.09
3.	57.0%	71.0%	50.0%	33.0%	60.0%
4.	14.0%	14.0%	100.0%	67.0%	60.0%

III. Students whose functional test level is <u>two years lower</u> than their age level test level.

Difference between higher level scores and functional level scores.

	Test A	Test B	Test C	Test D	Test E
1.	7	5	5	5	6
2.	1.47	2.72	1.97	2.76	1.08
3.	29.0%	60.0%	40.0%	60.0%	40.0%
4.	29.0%	20.0%	40.0%	60.0%	60.0%

greater than positive two or less than negative two. If the z- or t-score is greater than absolute two, the chances are high that the distribution of higher or lower level scores is different from the distribution of scores on the functional test level. The fourth line on the table contains the percent of negative z's or t's. Do the scores on a higher or lower test level go up or down from the scores on the functional test level? When there is a zero in line four, the scores went up for all ages. If there is a 50% in line four, in half of the age groups the scores went up and in half they went down. A high percent in line four on the sub-tables marked B (the difference between lower levels and functional test levels) indicates that in most age levels the scores went down on the lower test level.

Which test has the scores that change least when a different test level is administered to the same population of students? Sub-category median z-scores range between 1.04 and 3.05 on Test A, between 1.98 and 3.64 on Test B, between 1.17 and 2.84 on Test C, between 1.55 and 3.70 on Test D, and between 0.57 and 2.19 on Test E. Using these ranges as a rough guideline, we see the tests rank in the following order: Test E scores change the least, Tests A and C are about equal following Test E, and on the whole, Test B and Test D scores change more than the others. When we look at Table 5 to the t-scores between non-adjacent test levels (higher vs. lower) the ordering of the tests is about the same. These overall generalizations do not necessarily hold for every age group in every set of schools.

The data in Table 5 bring up several further questions about the effects of different levels of tests on students' scores. "Is one group of students any more affected by different levels of a test than another group of students?" The data show that testing up a higher level

affects the scores of students from Q3 and Q2 schools more than students from Q1 schools, and the students from Q4 schools are affected the least. There is not such a strong pattern when we look to the effects of testing down one level, and we cannot make any such general statement. However, on Test E, the least changeable test, Q3 and Q2 students have scores that change more than Q4 students' when they are tested down a level. Overall, on the three least changing tests, Test E, Test A, and Test C, the Q4 students are least affected by different test levels, the Q1 students next least affected, and the remaining half of the students (from Q3 and Q2) are affected the most, whether the levels go up or down.

Another question that we must ask of this data is, "Does it make more difference in students' scores if a test level is administered that is higher than the students' functional level than if it is lower than the students' functional level?" By investigating the median z-scores in Table 5, one can see that the Q4 students are more affected by testing down than by testing up, and that the Q3 and Q2 students are about equally affected. We cannot answer this question for Q1 students, since they were not tested down.

Up to this point we have been discussing absolute change in test scores, regardless of whether the scores went up on higher levels, and down on lower levels, as we often assume will be the case. The information presented in Table 5 indicates that students whose functional test level is equal to their age-level (students from Q-4 schools) are more likely to have test scores get higher on higher test levels and to get lower on lower test levels than any other students in the study. The students whose functional test level is one or two years lower than their age level test are much more likely to obtain lower scores on the higher test level and higher scores on the lower test levels. The tests themselves

differ on how many groups of students receive scores in the unexpected direction. Test A and Test B are least likely to give scores that are lower on a higher test level and higher on a lower test level than on the other tests. It is not necessarily desirable to have scores always go up when the test level is raised. For example, if students are placed on tests where they do not belong -- that is, where the testing is unreliable, the higher scores will have no educational usefulness.

The findings related to how test scores change when test levels are changed can be summarized as follows. Overall, scores on Test E change less from level to level than scores on any other of the tests. Test A and Test C scores change somewhat more than Test E scores, and scores on Test B and Test D change more than any other test scores. Students whose functional test level is equal to their age level test level (from Q4 schools) are the least affected by changing test levels (probably because testing is more reliable here), but testing down one level has a greater effect on mean scores than testing up one level. Students whose functional test level is one year lower than their age-level test level are affected more than other students by changing test levels. These students are equally affected by testing down one level as by testing up one level. The students whose functional test level is two years behind their age level test level, and were only tested up in the study, were affected more by testing up than the students from Q4 schools and less than students from Q3 and Q2 schools.

Summary and Conclusion

We have considered two important psychometric issues in this study, test reliability and the stability of scores from one test level to another. These two issues are related to each other in both theory

and practice, for a test that gives a different score depending on the level of the test cannot be reliable. We do not know which score provides a more accurate representation of students' achievement levels. These issues speak to our basic question of which test provides the most trustworthy scores -- the test that is most reliable and least likely to change if the test level is changed.

In addition to reliability and changeability of scores there are other very important measures of a test's suitability for a school system. The foremost is content validity of the test -- whether the test measures the same objectives and the same curriculum that we instruct. Content validity is the variable that should determine the selection of a standardized achievement test. When content validity is established, then we should look to reliability and decide whether a test is reliable or not.

We have found Test E to have the highest reliability coefficients in this study and also to display the least amount of change in scores from one test level to another. If Test E has the greatest content validity, then it would be an appropriate choice for the Chicago Board of Education.

Test A is also highly reliable and the scores change relatively little from one level to another. If its content validity is high, then Test A would also be appropriate for Chicago.

Test B is as reliable in Chicago as Test A. If it is found to have high content validity it may be suitable for Chicago, but there should be further study of the effect that different test levels will have on students' scores, for Test B was found in this study to change scores greatly from one test level to another.

Test C and Test D are less reliable than the other tests, probably because these two tests differ from the others in that they both have several levels of the test that are designed to test more than one age-level of students. Unless more accurate means of placement of students on to appropriate test levels on Test C and Test D can be devised, these tests do not appear to be as suitable for Chicago as the other three tests.

References

Agrawal, K. The effect of test level of the Iowa Tests of Basic Skills on grade equivalent scores. Department of Research and Evaluation, Chicago Board of Education, Technical Paper No. 2, August, 1978.

Ayrer, J. and McNamara, T. Survey testing on an out-of-level basis. Journal of Educational Measurement, 1973, 10, 79-83.

Iwanicki, L. A new generation of standardized achievement test batteries: a profile of their major features. Journal of Educational Measurement, 1980, 17, 155-162.

Long, J., Schaffran, J., and Kellogg, T. Effects of out-of-level testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977, 14, 203-213.

Thorndike, R.L. Educational measurement (second edition). Washington, D.C.: American Council on Education, 1971.

Wick, J. and Ward, F. Testing students at functioning reading level: A two-year report from Chicago Department of Research and Evaluation, Chicago Board of Education, 1977.

Yoshida, R.K. Out-of-level testing of special education students with a standardized achievement battery. Journal of Educational Measurement, 1976, 13, 215-221.

Appendix A

Description of Schools in the Study

The following table shows the average reading index score, and the average sizes and the locations of the schools that were assigned to each of the five tests in this study. It is readily apparent that the schools are nearly identical in terms of the average reading achievement, as measured by the 1980 citywide testing program. The average sizes of the schools vary somewhat more, but are approximately equal. The geographical distribution of schools is mostly even between the north and south sides of the city, although the STEP test was piloted in more south side schools than others. These minor variations in size and geographical location are compensated for by the equivalence on achievement in the five sets of schools.

Description of Schools

	Average Reading Index	Average Enrollment	Number Schools Dist. 1-10	Number Schools Dist. 11-12
CAT	250.2	740.7	6	6
ITBS	250.3	775.1	7	5
METRO	249.8	825.9	6	6
SRA	250.2	747.8	7	5
STEP	249.9	789.8	3	9